SCIENTIFIC OASIS

Journal of Operations Intelligence

Journal homepage: www.jopi-journal.org
ISSN: 3009-4267

Comparison of Machine Learning Approaches for Detecting COVID-19-Lockdown-Related Discussions During Recovery and Lockdown Periods

Mohammed Rashad Baker[1], A.H. Alamoodi[2,*], O.S. Albahri[3,4], A.S. Albahri[5,6], Salem Garfan[2], Amneh Alamleh[7], Moceheb Lazam Shuwandy[8], Ibrahim Alshakhatreh[9]

1 Software Department, College of Computer Science and Information Technology, University of Kirkuk, Kirkuk, Iraq
2 Faculty of Computing and Meta-Technology (FKMT), Universiti Pendidikan Sultan Idris (UPSI), Perak, Malaysia
3 Victorian Institute of Technology, Australia
4 Computer Techniques Engineering Department, Mazaya University College, Nasiriyah, Iraq
5 Department of Computer Technology Engineering, College of Information Technology, Imam Ja'afar Al-Sadiq University, Baghdad, Iraq
6 Iraqi Commission for Computers and Informatics (ICCI), Baghdad, Iraq
7 Department of Artificial Intelligence, Faculty of Information Technology, Zarqa University, Zarqa, Jordan
8 Computer Science Department, College of Computer Science and Mathematics, Tikrit University (TU), Tikrit, Iraq
9 College of management, department of business administration, National Yunlin University of Science and Technology, Yunlin, Taiwan

ARTICLE INFO

Article history:
Received 10 October 2023
Received in revised form 19 October 2023
Accepted 22 October 2023
Available online 24 October 2023

Keywords: COVID-19; Sentiment analysis; Lockdown; Machine Learning; Data balancing; SMOTE

ABSTRACT

Ever since COVID-19 was declared a pandemic, governments around the world have implemented numerous phases of lockdown measures to curb the spread of the virus. These lockdown tactics manifest themselves in the form of widespread fear and panic driven by social media discussions. Given that individuals hold diverse opinions about these lockdown measures during and after their completion, positive and negative lockdown-related discussions should be differentiated to further understand the major related issues and to make appropriate messaging and policy choices in the future. We conduct a sentiment analysis (SA) of COVID-19 lockdown-related tweets by using different machine learning (ML) classifiers and then evaluate their performance before and after using the synthetic minority oversampling technique (SMOTE). This research is performed in five phases, starting with data collection, followed by pre-processing the dataset, preparing the dataset by annotation, applying SMOTE, and using ML classifiers. We observe an improvement in accuracy (Acc), as confirmed by the Matthews correlation coefficient (MCC), across most classifiers, except for the k-nearest neighbour (KNN), whose Acc decreased from 0.82 to 0.59 and MCC decreased from 0.544 to 0.279 before and after SMOTE was applied. Despite the potential of SMOTE with some classifiers, this technique cannot be considered an ultimate solution, especially with other classifiers and datasets. The study provides insights into the need to evaluate and benchmark the integration of data balancing approaches with ML classifiers, in addition to considering additional metrics, such as MCC, for binary classification problems, especially in SA.

* *Corresponding author.*
E-mail address: mkonur@dho.edu.tr

https://doi.org/10.31181/jopi1120233

1. Introduction

Since its emergence in late 2019, the whole world has been suffering from the critical COVID-19 health crisis [1]. Immediately after this disease was reported in Wuhan, China, the World Health Organization (*WHO*) declared COVID-19 a pandemic as the virus started to spread worldwide [2]. To control the spread of the virus, many countries adopted safety measures, such as encouraging people to wear facemasks and practice social distancing, but the virus spread at an unprecedented rate and resulted in millions of fatalities [3]. The pandemic not only had health-related effects but also negatively influenced many other aspects of life, including economic, educational and social aspects [3]. With the rapid spread of COVID-19, governments around the world needed to make fast and timely decisions to control the situation, such as requiring most daily activities to be conducted online [4] and imposing travel restrictions and lockdowns [5]. After a few months, some countries relaxed their lockdown restrictions, but as the number of infections increased again, governments started to implement further restrictions, such as second and third lockdown measures, between 2020 and 2022 [6]. Aside from the economic and political impacts of these continuous lockdowns, how public opinion about these measures triggered a shift in public opinion during the lockdown and recovery periods also warrants investigation. Many people have flocked to online platforms, such as Twitter, Reddit and Facebook, to express their feelings and opinions about the topic [7], which resulted in millions of opinions and COVID-19-related discussions posted online daily [8]. Given the importance of public sentiment and the corresponding discussions related to lockdown measures, understanding the COVID-19-related discussions during the quarantine and recovery periods has become critical because the discussions across different periods may significantly vary. Public opinion about specific issues is influenced by the mental status of individuals, which, especially during the lockdown, can be influenced by frustration and depression [9]. These opinions may even shift during relaxed times even if the topic has not been changed. Consequently, an effective approach that can differentiate positive from negative public opinions about the same topic needs to be developed. To this end, many researchers have utilised computer science tools to process the language and text in social media posts. Amongst these tools, natural language processing (*NLP*) has been widely utilised in sentiment analysis (*SA*) [10].

SA or opinion mining is an NLP technique that has become increasingly important in analysing and processing large amounts of natural language data by establishing some principles through which individual opinions are analysed [11]. SA uses linguistics and AI to enable computers to understand human opinions, emotions and feelings about a particular subject are extracted from their own social media interactions [12]. As an emerging research field, Alamoodi, et al. [13] categorised *SA* research into three areas, namely, 1) lexicon-based, 2) machine learning (ML)-based and 3) hybrid SA research. SA has also been applied in a variety of applications, including scientific, social and commercial applications [14], and has even been used in analysing trending topics, such as the COVID-19 pandemic [13], patients attitude [15], vaccine hesitancy [16], and tracking online discourse [17]. To establish SA research, information regarding SA cases is collected from social media platforms such as Twitter, which is considered a valuable source of such information. Twitter has become an extremely popular platform on which users can freely express their feelings and share information daily about any event, including the pandemic. Using SA for Twitter can generate better insights about different case studies, offer a competitive advantage and provide optimal decisions for governments, stakeholders or organisations. Samuel, et al. [18] Used SA to gain insights into the progress of fear sentiment during the COVID-19 pandemic. Ghasiya and Okamura [19] Analysed the emergent and widely reported COVID-19-related topics in four countries and the associated sentiments. Obiedat, et al. [20] Used SA to help governments make effective decisions during the pandemic. Cotfas, et al. [21] Used SA to study the dynamics of COVID-19 vaccination opinions after the wide availability of vaccines has been announced. Mujahid, et al. [22] Used SA to study public opinion about online education during the pandemic. These studies constitute only a small portion of the literature that has utilised SA to explore various topics associated with COVID-19. Whilst the importance and implications of SA research focusing on COVID-19 are endless, some related areas and topics still require further investigation.

One major challenge for this work is to distinguish the public opinions and discussions about the quarantine and lockdown measures imposed by governments around the world. Specifically, the same person may share different opinions about the same topic across different periods. For instance, a person may publish a negative tweet about lockdown measures after s/he was forced to stay home, which eventually resulted in him/her losing his/her job and left him/her in a state of depression. However, after a certain period, this same person started publishing positive tweets about these lockdown measures after s/he was allowed again to work. Differentiating the opinions published across various periods can help governments understand the public sentiment towards their lockdown measures, which in turn will inform their future policies and decision-making strategies. Therefore, developing an SA text classification approach that can differentiate the COVID-19-lockdown-related discussions posted during lockdown or recovery periods becomes critical. One of the first tasks required in this process is to develop and compare different text classification approaches that can verify whether COVID-19-lockdown-related tweets have been posted during or after lockdown periods. To fulfil such task, a large amount of data needs to be processed by ML classifiers to build models with minimal human intervention. Whilst some previous works [3, 23-26] have attempted to compare the performance of different classifiers, only few have considered the presence of large-scale unbalanced datasets and the optimisation of parameters for classifier performance. This study aims to bridge this gap by answering the following question:

Q1: To which extent can ML classifiers distinguish tweets that are posted during lockdown and recovery periods?

Q2: To which extent can dataset balancing issues affect the performance of ML classifiers?

Q3: What suggestions can be offered to future SA research related to the evaluation and benchmarking of dataset balancing approaches?

This work aims to investigate the public sentiment regarding lockdowns and to determine whether discussions have been posted during or after lockdowns. This work also aims to highlight significant differences in the performance of ML classifiers before and after data balancing and propose some directions for future research regarding the evaluation and benchmarking of data balancing approaches used in SA research. This study sheds light on the general public sentiment by using a new, unexplored dataset. The goals of this research can be summarised as follows:

- To investigate the performance of ML classifiers in detecting lockdown-related discussions during or after lockdowns.
- To identify how dataset issues can affect the performance of ML classifiers for the presented case study; and
- To determine future potentials for evaluating and benchmarking dataset balancing approaches.

The rest of this paper is organised as follows. Section 2**.** reviews the literature. Section 3 describes the methodology. Section 4 presents the results and discussion. Section 5 concludes the paper.

2. Literature Review

This section discusses SA techniques, Twitter, related works and classification methods in four sub-sections.

2.1 Sentiment Analysis Techniques

As one of the most popular topics associated with NLP, SA uses the natural language toolkit (NLTK) implemented in Python with ML models to classify a particular target (binary or multi) depending on the case study and problem at hand. During this process, SA uses of variety of features and settings, such as term frequency–inverse document frequency (TF-IDF), bag of words, and N-gram [27]. In SA classification research, a supervised approach can be applied where ML classifiers are used for the classification based on fully labelled data [2]. In another hybrid approach, ML classifiers use where partially labelled training data to classify unlabelled data based on lexical methods [28]. SA research has mostly focused on social media platforms, such as Facebook, Twitter, Instagram and Reddit, given the wide availability of data [29]. During the COVID-19 pandemic, large volumes of news and discussions spread across these platforms, especially on Twitter, as it provides a strong platform for people to share their opinions in a timely manner. Therefore, using Twitter for SA is especially suitable for the case study presented in this research.

2.2 Twitter for Sentiment Analysis Techniques

Twitter is amongst the most widely used social media platforms in the world. According to Abraham, et al. [30], Twitter has 330 million monthly active users (out of 1.3 billion registered accounts) and publishes over 500 million tweets each day. Therefore, Twitter is not only a comprehensive data source for analysing public sentiment about almost any topic but also a valuable data asset for academic researchers to develop predictive models on large-scale real-time data, which is tremendously helpful during emergencies and crises, such as the COVID-19 pandemic. One of the most well-known cases where Twitter data were used for emergencies was reported in the study of Sakaki, et al. [31], who used Twitter data to develop an event detection system for identifying urgent events, including earthquakes. Other applications include Chew and Eysenbach [32], who used Twitter data to understand public sentiment during the pandemic, and Jain and Kumar [33], who used Twitter data to track the Influenza-A pandemic in India. Aside from health, SA has been widely adopted in other areas, including tourism [34], business [35] and education [36]. In this study, SA is mainly associated with COVID-19-related discussions during lockdown and recovery periods. COVID-19-related studies that fall within the scope of this research are reviewed accordingly in the following sub-section.

2.3 Related Works

SA is an NLP application that analyses public sentiment and perceptions about a certain topic by using large amounts of unstructured data. SA has been actively used in COVID-19 research since the beginning of the pandemic. This section then reviews the use of SA in the related literature. Al-Hashedi, et al. [37] Used SA, Word2Vec features and 6 ML models to study COVID-19-related conspiracy theories on Arabic Twitter. They found that the performance of ML models (single and ensemble) can be enhanced by using data balancing techniques, such as the synthetic minority oversampling technique for nominal and continuous. Samuel, et al. [18] Studied the progress of fear sentiment over time by focusing on COVID-19-specific tweets and using ML models whilst considering their performance over tweets with varying lengths. They found that the Naïve Bayes (NB) classifier performed very well for short tweets with 91% accuracy compared with logistic regression (LR), which only obtained 74% accuracy. The performance of both these classifiers degraded along with an increasing tweet length. Misinformation not only spreads false news about the pandemic but also in evokes panic and fear amongst people, thereby negatively affecting culture, economics and healthcare. To address this problem, Alenezi and Alqenaei [38] established an effective ML model for the detection of misinformation regarding COVID-19 and obtained much more superior results compared with the literature by using long short-term memory (LSTM), multichannel convolutional neural network and k-nearest neighbours (KNN). In the same vein, Al-Ahmad, et al. [1] detected COVID-19-related misinformation by reducing the number of symmetrical features after applying particle swarm optimisation, genetic algorithm and salp swarm algorithm. This approach reported a better accuracy and outperformed the other classifiers in the literature. Yao, et al. [2] Used ML models to investigate public sentiment toward the pandemic in different mega-cities, such as New York, Los Angeles and London. They found that majority of COVID-19-related sentiments was posted at the beginning of 2020 (around mid-March) and on early May 2020 and suggested that public sentiment is more sensitive to quarantine orders compared with reported statistics. Anti-vaccination attitudes, which were prevalent not only during the COVID-19 pandemic but also in other pandemics reported in history, have been identified as one of leading factors that contribute to vaccine hesitancy. In response to the proliferation of anti-vaccination content on social media, To, et al. [23] evaluated the performance of different NLP models in identifying anti-vaccination tweets published during the COVID-19 pandemic by using bidirectional encoder representations from transformers (BERT) and the bidirectional long short-term memory networks with pre-trained GLoVe embeddings along with some classic ML models, including support vector machine (SVM) and NB. They found that BERT models achieved excellent performance in identifying anti-vaccination content and can thus be used in future studies. Alabrah, et al. [39] Used ML-based models and three sentiment extraction approaches (i.e. Ratio, TextBlob and VADER) to analyse COVID-19-vaccine-related discourses on social media in Gulf Cooperation Council countries. The latter scores results when used with KNN and Ensemble boost presented 94.01% classification accuracy, and the proposed work deemed robust in classifying and determining sentiments in Twitter discourse. Cotfas, et al. [21] Analysed the dynamics of COVID-19-vaccination-related opinions a month after the announcement of vaccine availability. They utilised both ML and deep learning models to compare their performance. They found that positive tweets outnumbered the negative ones and proposed that both ML and deep learning approaches can help governments formulate an appropriate communication strategy. Aljabri, et al. [24] Explored public sentiment and acceptance of distance learning measures by using ML models along with different features and extraction techniques. They found that LR obtained the best accuracy of 0.899 and can therefore help governments around the world in improving their distance learning systems. Rustam, et al. [25] Compared the performance of supervised ML models in the multi-label classification of the sentiment presented in COVID-19-related tweets by using different feature sets. With the highest accuracy of 0.93, trees classifiers can facilitate the identification of COVID-19-related sentiments on social media and subsequently contribute to the making of informed decisions for handling the pandemic. Aljameel, et al. [3] developed ML models with different feature sets to predict public awareness about COVID-19 precautionary measures. Amongst these models, the SVM ML classifier with bigram in TF-IDF obtained the highest accuracy of 85%. Gulati, et al. [26] compared the performance of different ML classifiers by using a dataset of tweets related to the COVID-19 pandemic. By using 3 feature modes and 7 classifiers, they found that SVM obtained the highest accuracy amongst all ML models. Therefore, SVM may play a key role in predicting important information about the pandemic and its spread. As can be seen from the above review, the majority of the SA literature has focused on different cases, including misinformation, public awareness, vaccination and education. However, only few studies have investigated the performance of ML classifiers in identifying lockdown-related discussions during lockdown or recovery period. In addition, the performance of ML models before and after data balancing has received limited research attention, and a method for improving their performance needs to be developed. This study combines the benefits of Twitter data and machine learning methods on large-scale, high-column dataset to help future researchers in understanding public sentiment toward COVID-19 based on related discussions during two different periods.

2.4 Classification Methods

Many studies have used ML classifiers for text classification in SA, and each classifier has obtained different results based on the case study applied. This section discusses the main classifiers utilised in this work, including LR, random forest (RF), decision tree (DT), KNN, AdaBoost, XGBoost and artificial neural network (ANN).

2.4.1 Logistic Regression

LR is one of the most common ML classification approaches applied in the literature [40]. This classifier follows the probability concept for a single trial outcome by using a logistic function [41], whose outcome probability is either 1 or 0. Given that this classifier has been widely used in SA research [42], LR is included amongst the ML classifiers to be compared in this study.

2.4.2 Random Forest

RF is an ensemble classifier where multiple decision trees are trained in parallel via bootstrapping followed by bagging [43]. RF is known for its good performance in addressing classification problems and does not require feature scaling. This classifier has been widely used in SA classification [44], and its parameters can be easily adjusted compared with other classifiers [45]. Given these benefits, RF is included amongst the classifiers compared in this work.

2.4.3 Decision Tree

DT is an ML classifier that operates a hierarchical tree, uses attribute value conditions to split the training data into smaller parts and performs different tests to show the tree branches. Each branch slops from the node matches to the feature value. To classify an instance, the feature of the parent node is checked, and then the branch of the tree is examined to determine the value of the feature for a particular instance. DT has been widely used in addressing SA classification problems [46] and is therefore included in this research.

2.4.4 K-Nearest Neighbour

KNN is an important ML classifier that uses instance-based learning and similarity measurements for text classification purposes, where the similarity between two points are measured by estimating their distance, proximity or closeness function [47]. In KNN, the number of nearest neighbours can be either estimated or specified within a fixed point radius [48]. This classifier has been widely used in addressing text classification and SA problems [41] and is therefore examined in this work.

2.4.5 Naïve Bayes

Based on the Bayes theorem, NB is a simple ML approach [47] that applies maximum a posteriori estimation to identify the associated class, hence making this approach particularly effective in addressing text classification problems [49]. NB efficiently performs with a limited size of training data and is based on conditional and data shape distribution probabilities [40]. Accordingly, the performance of NB is compared with that of other classifiers in this work.

2.4.6 AdaBoost

AdaBoost is an ML algorithm that combines other weak classifiers for a better performance [50]. This algorithm works by calling a weak classifier several times and providing a different distribution over the training data in each call instance. AdaBoost can filter out unnecessary training data features and place them on key training data. This algorithm has been used in SA research [51] and is considered a good fit for the performance comparison in this work.

2.4.7 XgBoost

Extreme gradient boosting, also known as XgBoost, is a quick and scalable ML classifier based on the gradient tree boosting technique that has been recognised for its remarkable performance in addressing many ML problems, including SA ones [52]. This classifier works in parallel with high scalability and performance and applies ensemble methods in its boosting part [53]. With its parameters, this classifier can be tuned differently to yield good results and hence was included in this study for the performance comparison.

2.4.8 Multi-Layer Perceptron

Multi-layer perceptron (MLP) is an ANN classifier known for its considerable performance in addressing a variety of SA and ML classification problems [54]. MLP often uses at least three layers to build relationships between the input and output layers where neurons with variable weights make up each layer [55]. In MLP, every neuron in the next layer receives input from the input nodes or neurons (i.e. hidden layers) of the previous layer. The neurons in the hidden layer are linked to one another [56]. The above ML classifiers yield excellent performance in a variety of cases, whether in SA or other ML context problems. Sometimes, a classifier for an SA problem can yield a different performance if executed on two different datasets or case studies. The performance of these classifiers is influenced by several factors, including dataset length, classification target, number of classes to be predicted/classified, features and parameters. Therefore, there is no definitive way to select a particular classifier for a certain case—unless this classifier shows a good performance from the start or at least compared with others—to confirm whether this classifier is the best fit for such case. To this end, a total of n=8 classifiers are compared in this work to determine which of them best operates on the SA problem presented in this research. Performance evaluation matrices will be used to determine the overall performance of each model. These matrices include accuracy, recall, precision, F-score and Matthews's correlation coefficient (MCC), which are computed from the confusion matrix of each classifier. Additional information on the definition and equation for the first four matrices and MCC can be found in Han, et al. [57] and Chicco, et al. [58], respectively. Many studies have used ML classifiers for text classification in SA, and each classifier has obtained different results based on the case study applied. This section discusses the main classifiers utilised in this work, including LR, random forest (RF), decision tree (DT), KNN, AdaBoost, XGBoost and artificial neural network (ANN).

3. Methodology

This section discusses the methodological phases adopted in this research. As shown in Figure 1, this study has four main phases. The first phase discusses the data collection procedure, starting from the identification of keywords to the assembly of the entire dataset. The second phase discusses the pre-processing steps applied on the assembled dataset, starting from the filtration until the end of the processing. The third phase prepares the dataset for the ML classification with tasks such as annotation procedure, balancing and feature extraction. The fourth phase takes the output of the previous phase, applies the ML predictive models proposed in this research and discusses their performance.

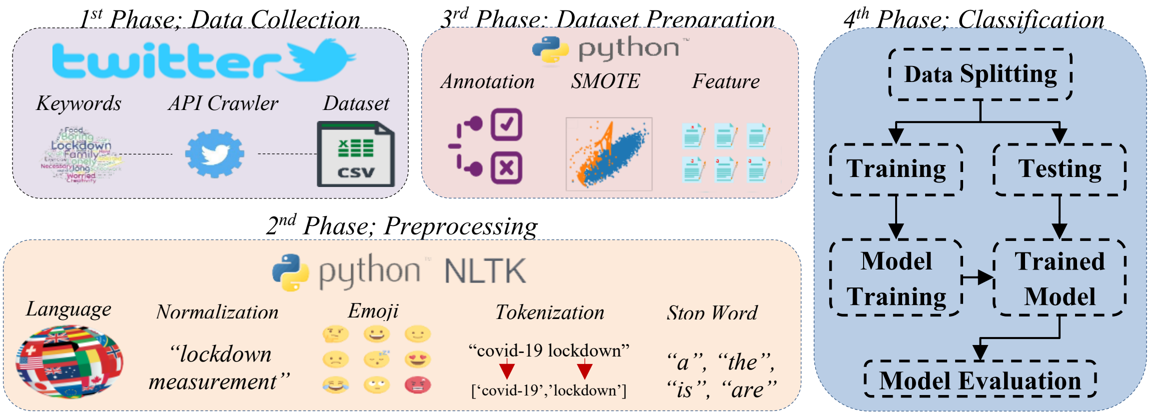

Fig. 1. Methodology Phases

3.1 Data Collection

The data were collected starting from the announcement of the first COVID-19 lockdown in Malaysia (March 2020) up to November 2021. Between these dates, a total of three full lockdowns were implemented in the country with different recovery periods in between. The data collection was aimed at Twitter as the most used platform for sharing opinions and tweets regarding the quarantine and lockdown. The lockdown phases in Malaysia are known as movement control orders (MCO). Therefore, the following keywords were used to fetch the lockdown-related tweets: #Movement_Control_Order, #MCO, #MCO 2.0, #MCO 3.0, #COVID-19_quarantine, #COVID-19_Lockdown and #Lockdown. These keywords fetched more than (n=3,000,000) tweets using the Twint project tool, which allows users to scrape tweets and other historical information [59]. All the collected tweets after discarding the duplicated and unrelated ones were exported as comma-separated values (.csv) for the analysis.

3.2 Pre-Processing

Pre-processing is an essential step prior the actual sentiment analysis using ML classifiers. The data collected from Twitter, in their original form, should undergo a filtration process for them to be compatible for further processing. To this end, different Python tools and libraries were used, including the Pandas-Python Data Analysis Library and Python NLP toolkit. The first step in the pre-processing process is normalisation, which cleans the collected data. This process involves filtering out non-English text, emojis, URLs and special characters from the data, converting all upper case letters into small letters to reduce the complexity of using text in the analysis and tokenisation, wherein the text is converted into tokens separated by a white space. In tokenisation, all special characters are removed, the word boundaries are determined, and the abbreviations and numbers are processed. In the last step, the stop words are removed to eliminate those words that do not add much information to the text about a particular topic. Examples of stop words in the English language are 'is', 'a', 'an' and 'the'.

3.3 Annotation

In annotation, all tweets and their categories are labelled in accordance with the research main case for the classification [60]. Accordingly, the dataset was annotated into two categories, namely, lockdown and recovery, for the binary classification. Lockdown includes the COVID-19-related discussions about the lockdown and quarantine measures implemented during the lockdown period, whereas recovery includes the corresponding discussions during the recovery period. A two-step filtration process was then implemented, in which the first step selected all tweets with at least one keyword related to COVID-19 lockdown discussions in Malaysia, and the second step used the publication date of each tweet to determine whether this tweet was posted during the lockdown or recovery period. After this process, a total of (n=1,050,803) tweets were retained for the analysis. The distribution of these tweets between the two categories is shown in Fig. **1**.

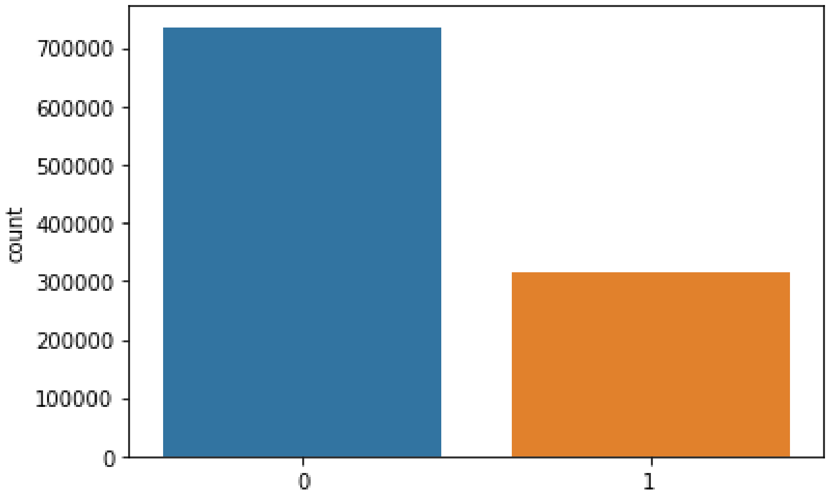

Fig. 1. Categories Distribution in the Dataset

As shown in Figure 2, the lockdown category included (n=316,108) tweets (labelled as 1), whereas the recovery category included (n=734,695) tweets (labelled as 0), with the lockdown tweets clearly outnumbering the recovery ones. This distribution can affect the performance of the ML classifiers in accurately classifying the problem of the case study. To address this issue, the synthetic minority oversampling technique (SMOTE) was applied to balance the distribution between these categories. The results are presented before and after the application of SMOTE to highlight how much this technique improves the classification.

3.4 SMOTE

SMOTE is a data balancing technique used in ML research. In this approach, synthetic samples of minority classification labels are generated so that the number of samples from each group is almost the same [61]. As discussed in Section 0, the dataset did not have an equal representation of the annotated categories, which may lead to the overfitting of the ML models. The effects of SMOTE on the data ratio and on the performance of ML classifiers are discussed in Sections 0 and 0.

3.5 Feature Extraction

TF-IDF has been widely used in SA research to extract weighted features from the data and to assign each data term with a weight value to enhance the performance of ML classifiers [62]. TF-IDF focuses on the most distinctive words, hence allowing this approach to overcome the limitation of depending on word counts in SA research. The mathematical functions for TF-IDF are

$$\mathrm{tf(t,d)} = \log\left(1 + \mathrm{f_{t,d}}\right) \tag{1}$$

$$\mathrm{Idf(t)} = \log\left(\frac{1 + \mathrm{N}}{1 + \mathrm{n}_t}\right) \tag{2}$$

where $\mathrm{tf(t,d)}$ represents the count of term t in document d, N represents the total number of documents and n represents the number of documents containing the term t.

3.6 Data Splitting

The collected data were split into an 80:20 ratio, of which 80% were used for training the model, and the remaining 20% were used for testing the model. A 10-fold cross-validation was then performed. Prior this process, the data were shuffled to enhance the generalisability of the classification results, reduce the variance and avoid model overfitting.

3.7 Performance Metrics

The performance evaluation measures used in this research include accuracy (Acc), Precision (Pr), Recall (Re), F1 score and MCC, which are defined as follows:

$$Accuracy(Acc) = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision\ (Pr.) = \frac{TP}{TP + FP} \tag{4}$$

$$Recall\ (Re.) = \frac{TP}{TP + FN} \tag{5}$$

$$F1\ Score = 2x\ \frac{Pr. x\ Re.}{Pr. + Re.} \tag{6}$$

$$MCC = \frac{TP\ x\ TN - FP\ x\ FN}{\sqrt{(TP + FP)x\ (TP + FN)x(TN + FP)x(TN + FN)}} \tag{5}$$

where TP, FP, TN and FN denote true positive, false positive, true negative and false positive, respectively.

4. Result

Two experiments were conducted on all classifiers. The data were imbalanced in the first experiment, and this imbalance issue was resolved using SMOTE in the second experiment.

4.1 Result without SMOTE

All classifiers in the original dataset with a class imbalance issue were used in the first experiment. The results are reported in

Table 1
Results without Using SMOTE

Model	Accuracy	Class	Precision	Recall	F-Score	MCC
AdaBoost	0.78	0	0.77	0.97	0.86	0.418
		1	0.81	0.34	0.47	
		Macro avg	0.79	0.65	0.67	
DT	0.79	0	0.83	0.87	0.85	0.479
		1	0.66	0.60	0.63	
		Macro avg	0.75	0.73	0.74	
KNN	0.82	0	0.82	0.95	0.88	0.544
		1	0.81	0.52	0.63	
		Macro avg	0.82	0.73	0.76	
LR	0.82	0	0.88	0.86	0.87	0.580
		1	0.69	0.73	0.71	
		Macro avg	0.79	0.79	0.79	
MLP	0.79	0	0.88	0.81	0.84	0.527
		1	0.62	0.75	0.68	
		Macro avg	0.75	0.78	0.76	
NB	0.81	0	0.81	0.95	0.87	0.513
		1	0.81	0.47	0.60	
		Macro avg	0.81	0.71	0.74	
RF	0.75	0	0.73	1.00	0.85	0.331
		1	0.97	0.16	0.27	
		Macro avg	0.85	0.58	0.56	
XgBoost	0.81	0	0.79	0.98	0.88	0.513
		1	0.89	0.41	0.56	
		Macro avg	0.84	0.69	0.72	

KNN and LR outperformed all other classifiers in terms of accuracy (0.82), followed by XgBoost and NB (0.81), DT and MLP (0.79), AdaBoost (0.78) and RF (0.75). Apart from accuracy, MCC is considered a more robust performance evaluation approach for binary classification problems, especially in the case of balanced and unbalanced datasets. LR obtained the highest MCC (0.580), followed by KNN (0.544), MLP (0.527) and RF (0.331).

4.2 Result with SMOTE

In the second experiment, the same classifiers were applied after using SMOTE to balance the dataset distribution. This experiment aimed to show how SMOTE can enhance the performance of classifiers when applied on an unbalanced dataset. The results are shown in

Table 2
Results After Using SMOTE

Model	Accuracy	Class	Precision	Recall	F-Score	MCC
AdaBoost	0.70	0	0.66	0.83	0.74	0.423
		1	0.77	0.58	0.66	
		Macro avg	0.72	0.70	0.70	
DT	0.83	0	0.82	0.84	0.83	0.655
		1	0.84	0.81	0.82	
		Macro avg	0.83	0.83	0.83	
KNN	0.59	0	0.88	0.21	0.34	0.279
		1	0.55	0.97	0.70	
		Macro avg	0.71	0.59	0.52	
LR	0.85	0	0.85	0.86	0.85	0.700
		1	0.85	0.84	0.85	
		Macro avg	0.85	0.85	0.85	
MLP	0.85	0	0.88	0.82	0.85	0.711
		1	0.83	0.89	0.86	
		Macro avg	0.86	0.85	0.85	
NB	0.80	0	0.73	0.93	0.82	0.618
		1	0.91	0.66	0.77	
		Macro avg	0.82	0.80	0.79	
RF	0.71	0	0.65	0.92	0.76	0.456
		1	0.86	0.49	0.63	
		Macro avg	0.75	0.71	0.69	
XgBoost	0.81	0	0.74	0.95	0.83	0.639
		1	0.93	0.66	0.77	
		Macro avg	0.83	0.81	0.80	

In this experiment, MLP and LR obtained the highest accuracy (0.85), followed by DT (0.83), XgBoost (0.81), NB (0.80), RF (0.71) and AdaBoost (0.70). In terms of MCO, MLP showed the best performance (0.711), followed by LR (0.700), DT (0.655) and AdaBoost (0.423). These results clearly show that the application of SMOTE improved the accuracy of some classifiers whilst degrading the performance of others. Nevertheless, the application of this technique improved the MCO of all classifiers, hence proving the suitability of SMOTE in performance evaluation after balancing the dataset. The following section discusses the improvements and degradations in the performance of the classifiers before and after the application of SMOTE.

4.3 Comparative Analysis

This section compares the performance of all classifiers before and after the application of SMOTE based on their accuracy and MCC values. The comparison is illustrated on Fig. 2.

Fig. 2. Comparative Analysis

As shown in Figure 3, the application of SMOTE improved the accuracy of three classifiers, namely, DT (from 0.78 to 0.83), LR (from 0.82 to 0.85) and MLP (from 0.79 to 0.85). XgBoost showed no changes in its accuracy (0.81) before and after the application of SMOTE, whereas Adaboost, KNN, NB and RF reported a decrease in their accuracy, hence suggesting that SMOTE is not suitable for improving the accuracy of classifiers. However, in terms of MCC, which has been proven to be more robust and reliable than balanced accuracy in over F1 score and accuracy in binary classification evaluation [58, 63]. Figure 3 shows that most classifiers (n=7) have improved their MCC after the application of SMOTE, with MLP showing the largest increase (18.4), followed by DT (17.6), XgBoost (12.6), RF (12.5), LR (12) and AdaBoost (0.5). Consistent with the results for accuracy, KNN reported a lower MCC after the application of SMOTE.

4.4 Discussion and Synthesis

The majority of the ML classifiers compared in this research has shown improvements in their performance as demonstrated in their MCC scores, which has been demonstrated as one of the most suitable approaches for addressing classification problems, especially when the data are balanced using artificial means, such as SMOTE. Only the KNN classifier showed no improvements in both its MCC and accuracy. Surprisingly, KNN demonstrated a better performance in the presence of largely unbalanced data, thereby suggesting that in spite of the potential of data balancing approaches (or at least SMOTE), their utilisation in balancing a dataset may not always be applicable across all ML classifiers. Therefore, the performance of these different classifiers should be measured by using other data balancing approaches.

4.5 Potential Future Research Directions

This section presents some promising directions for future COVID-19 SA research in terms of evaluating and benchmarking data balancing techniques. Results of this study show that data balancing techniques are primarily used to rectify unbalanced datasets and to ensure that ML and other predictive analytic classifiers can perform accurately. Although many data balancing methods are available, not all of them can improve the performance of classifiers. For instance, SMOTE negatively influenced the performance of the KNN classifier as can be seen in its MCC and accuracy. Whilst other data balancing methods can be used, their positive effects on the performance of classifiers cannot be guaranteed. Meanwhile, determining the most suitable classifier-data balancing combination for SA research is a complex process that is affected by several issues. Firstly, many ML classifiers and data balancing approaches are available for SA classification research. Secondly, the results of evaluation metrics for classification tasks may greatly vary from one metric to another. For example, a classifier with high accuracy may perform poorly for the other performance assessment metrics. Furthermore, the majority of comparative ML SA studies does not offer an ML-data balancing combination that is best suited for addressing the binary classification issue in COVID-19 research. Therefore, the evaluation and benchmarking of an ML-data balancing combination for the SA binary classification problem in COVID-19 research is a challenging multi-attribute decision-making problem that comes within MCDM. MCDM, which has been characterised as a 'decision theory extension that covers any choice with multiple objectives', is a technique for evaluating choices based on separate and sometimes contradictory criteria and merging these choices into a single overall evaluation. This approach involves several activities, including organising, planning and addressing various decision issues based on a variety of criteria. MCDM approaches often need decision makers to submit qualitative and/or quantitative evaluations to evaluate the performance of each alternative in relation to each criterion and the relative relevance of the evaluation criteria in relation to the overall aim. We plan to integrate MCDM alongside the evaluation and benchmarking issue of ML-data balancing combination in future SA research for the binary classification problem.

5. Conclusion

The COVID-19 pandemic has affected the world in many ways and even resulted in huge fatalities. As a result, the pandemic necessitated a quick and decisive action from the government to the extent that most daily operations were moved online, travel restrictions were enforced, and lockdowns were implemented on a regular basis. The reaction of the public to such decisions during periods of lockdown and recovery has generated millions of comments and discussions every day on social media. These discussions about the same topic should be differentiated from one another based on when they occur to understand the dominating public opinion during periods of lockdown or recovery. To this end, this study measured the online discussions related to the continuous lockdown measures by using SA with ML classifiers. A dataset of tweets from Malaysia was obtained by using the Twint tool and several keywords related to MCO. A variety of pre-processing steps were also applied on the dataset, including text normalisation and emoji and stop words removal. Afterwards, the data were annotated using Date to label tweets into binary classes to be used with the ML classification. The collected tweets had an unbalanced distribution, which may affect the performance of the classifiers. Therefore, a data balancing approach was used to obtain a balanced dataset. TD-IDF was used as feature engineering approach whose results were fed into (n=8) ML classifiers, namely, AdaBoost, DT, KNN, LR, MLP, NB, RF and XgBoost, whose results before and after the application of the SMOTE data balancing approach were compared. All these classifiers reported accuracies ranging from 0.75 to 0.82, with both KNN and LR obtaining the highest accuracy and RF obtaining the poorest accuracy (0.75). However, accuracy is not the best measure of classifier performance. For this reason, the MCC of each classifier was measured. LR obtained the highest MCC (0.580), whereas RF obtained the lowest MCC (0.331). After data balancing, MLP obtained the best accuracy (0.85) and highest MCC score (0.711) amongst all classifiers. Surprisingly, the application of SMOTE degraded the performance of some classifiers to their worst levels. For instance, the accuracy and MCC measurements of KNN reduced from 0.82 to 0.59 and from 0.544 to 0.279, respectively, thereby suggesting that data balancing techniques, or at least the ones utilised in this research, do not guarantee the performance of classifiers. Therefore, future COVID-19 SA research should evaluate and benchmark other approaches aside from data balancing with different ML classifiers and across different SA evaluation datasets. Future research may also extend this work by introducing deep learning classifiers, different feature settings, data balancing techniques and predictive analyses of the SA dataset presented in this study and other benchmarking datasets available in the literature. The findings of this work and its future extensions not only will contribute additional technical knowledge into ML, its settings and parameters but also help public health authorities make decisions in response to emergency situations.

Author Contributions

Conceptualization, methodology, software, validation, formal analysis, investigation, M.R.B., A.H.A., O.S.A., A.S.A., S.G., A.A., M.L.S., and I.A.; resources, data curation, A.A., M.L.S., and I.A.; writing—original draft preparation, writing—review and editing, A.A., M.L.S., and I.A.; supervision, O.S.A., and A.S.A. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

Data will be made available on request.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was not funded by any grant.

References

[1] Al-Ahmad, B., Al-Zoubi, A. M., Abu Khurma, R., & Aljarah, I. (2021). An evolutionary fake news detection method for COVID-19 pandemic information. *Symmetry*, *13*(6), 1091. https://doi.org/10.3390/sym13061091

[2] Yao, Z., Yang, J., Liu, J., Keith, M., & Guan, C. (2021). Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19. *Cities*, *116*, 103273.. https://doi.org/10.1016/j.cities.2021.103273

[3] Aljameel, S. S., Alabbad, D. A., Alzahrani, N. A., Alqarni, S. M., Alamoudi, F. A., Babili, L. M., ... & Alshamrani, F. M. (2021). A sentiment analysis approach to predict an individual's awareness of the

precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia. *International journal of environmental research and public health*, *18*(1), 218.. https://doi.org/10.3390/ijerph18010218

[4] Wahl, B., Cossy-Gantner, A., Germann, S., & Schwalbe, N. R. (2018). Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings?. *BMJ global health*, *3*(4), e000798.. http://dx.doi.org/10.1136/bmjgh-2018-000798

[5] Blakely, T., Thompson, J., Bablani, L., Andersen, P., Ouakrim, D. A., Carvalho, N., ... & Stevenson, M. (2021, July). Association of simulated COVID-19 policy responses for social restrictions and lockdowns with health-adjusted life-years and costs in Victoria, Australia. In *JAMA Health Forum* (Vol. 2, No. 7, pp. e211749-e211749). American Medical Association. https://doi.org/10.1001/jamahealthforum.2021.1749

[6] Blakely, T., Thompson, J., Bablani, L., Andersen, P., Ouakrim, D. A., Carvalho, N., ... & Stevenson, M. (2021, July). Association of simulated COVID-19 policy responses for social restrictions and lockdowns with health-adjusted life-years and costs in Victoria, Australia. In *JAMA Health Forum* (Vol. 2, No. 7, pp. e211749-e211749). American Medical Association. https://doi.org/10.1001/jamahealthforum.2021.1749

[7] Basile, V., Cauteruccio, F., & Terracina, G. (2021). How dramatic events can affect emotionality in social posting: The impact of COVID-19 on Reddit. *Future Internet*, *13*(2), 29. https://doi.org/10.3390/fi13020029

[8] Antonakaki, D., Fragopoulou, P., & Ioannidis, S. (2021). A survey of Twitter research: Data model, graph structure, sentiment analysis and attacks. *Expert Systems with Applications*, *164*, 114006. https://doi.org/10.1016/j.eswa.2020.114006

[9] Haque, M., Haque, I. E., Ziku, M. N. E. A., Ahamed, N., & Hossain, M. S. (2021). COVID-19 Pandemic and Its Effects on Youth Mental Health in Bangladesh. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, *6*(10), 365-377. https://doi.org/10.47405/mjssh.v6i10.1071

[10] I. Lauriola, A. Lavelli, and F. Aiolli, "An Introduction to Deep Learning in Natural Language Processing: Models, Techniques, and Tools," *Neurocomputing,* 2021. https://doi.org/10.1016/j.neucom.2021.05.103

[11] P. Tyagi and R. Tripathi, "A review towards the sentiment analysis techniques for the analysis of twitter data," in *Proceedings of 2nd international conference on advanced computing and software engineering (ICACSE)*, 2019. https://dx.doi.org/10.2139/ssrn.3349569

[12] Lauriola, I., Lavelli, A., & Aiolli, F. (2022). An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing*, *470*, 443-456.. https://doi.org/10.1007/s10115-018-1236-4

[13] Alamoodi, A. H., Zaidan, B. B., Zaidan, A. A., Albahri, O. S., Mohammed, K. I., Malik, R. Q., ... & Alaa, M. (2021). Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review. *Expert systems with applications*, *167*, 114155. https://doi.org/10.1016/j.eswa.2020.114155

[14] Keramatfar, A., & Amirkhani, H. (2019). Bibliometrics of sentiment analysis literature. *Journal of Information Science*, *45*(1), 3-15. https://doi.org/10.1177/0165551518761013

[15] Roccetti, M., Marfia, G., Salomoni, P., Prandi, C., Zagari, R. M., Kengni, F. L. G., ... & Montagnani, M. (2017). Attitudes of Crohn's disease patients: infodemiology case study and sentiment analysis of Facebook and Twitter posts. *JMIR public health and surveillance*, *3*(3), e7004. https://doi.org/10.2196/publichealth.7004

[16] Alamoodi, A. H., Zaidan, B. B., Al-Masawa, M., Taresh, S. M., Noman, S., Ahmaro, I. Y., ... & Salahaldin, A. (2021). Multi-perspectives systematic review on the applications of sentiment analysis for vaccine hesitancy. *Computers in Biology and Medicine*, *139*, 104957. https://doi.org/10.1016/j.compbiomed.2021.104957

[17] Jang, H., Rempel, E., Roth, D., Carenini, G., & Janjua, N. Z. (2021). Tracking COVID-19 discourse on twitter in North America: Infodemiology study using topic modeling and aspect-based sentiment analysis. *Journal of medical Internet research*, *23*(2), e25431. https://doi.org/10.2196/25431

[18] Samuel, J., Ali, G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314. https://doi.org/10.3390/info11060314

[19] Ghasiya, P., & Okamura, K. (2021). Investigating COVID-19 news across four nations: A topic modeling and sentiment analysis approach. *Ieee Access*, *9*, 36645-36656. https://doi.org/10.1109/ACCESS.2021.3062875

[20] Obiedat, R., Harfoushi, O., Qaddoura, R., Al-Qaisi, L., & Al-Zoubi, A. M. (2021). An evolutionary-based sentiment analysis approach for enhancing government decisions during COVID-19 pandemic: The case of jordan. *Applied Sciences*, *11*(19), 9080. https://doi.org/10.3390/app11199080

[21] Cotfas, L. A., Delcea, C., Roxin, I., Ioanăş, C., Gherai, D. S., & Tajariol, F. (2021). The longest month: analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *Ieee Access*, *9*, 33203-33223. https://doi.org/10.1109/ACCESS.2021.3059821

[22] Mujahid, M., Lee, E., Rustam, F., Washington, P. B., Ullah, S., Reshi, A. A., & Ashraf, I. (2021). Sentiment analysis and topic modeling on tweets about online education during COVID-19. *Applied Sciences*, *11*(18), 8438. https://doi.org/10.3390/app11188438

[23] To, Q. G., To, K. G., Huynh, V. A. N., Nguyen, N. T., Ngo, D. T., Alley, S. J., ... & Vandelanotte, C. (2021). Applying machine learning to identify anti-vaccination tweets during the COVID-19 pandemic. *International journal of environmental research and public health*, *18*(8), 4069. https://doi.org/10.3390/ijerph18084069

[24] Aljabri, M., Chrouf, S. M. B., Alzahrani, N. A., Alghamdi, L., Alfehaid, R., Alqarawi, R., ... & Alduhailan, N. (2021). Sentiment analysis of Arabic tweets regarding distance learning in Saudi Arabia during the COVID-19 pandemic. *Sensors*, *21*(16), 5431. https://doi.org/10.3390/s21165431

[25] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, *16*(2), e0245909. https://doi.org/10.1371/journal.pone.0245909

[26] Gulati, K., Kumar, S. S., Boddu, R. S. K., Sarvakar, K., Sharma, D. K., & Nomani, M. Z. M. (2022). Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic. *Materials Today: Proceedings*, *51*, 38-41. https://doi.org/10.1016/j.matpr.2021.04.364

[27] Shahana, P. H., & Omman, B. (2015). Evaluation of features on sentimental analysis. *Procedia Computer Science*, *46*, 1585-1592. https://doi.org/10.1016/j.procs.2015.02.088

[28] Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv preprint arXiv:1103.2903*. https://doi.org/10.48550/arXiv.1103.2903

[29] Ali, K., Dong, H., Bouguettaya, A., Erradi, A., & Hadjidj, R. (2017, June). Sentiment analysis as a service: a social media based sentiment analysis framework. In *2017 IEEE international conference on web services (ICWS)* (pp. 660-667). IEEE. https://doi.org/10.1109/ICWS.2017.79

[30] Abraham, J., Higdon, D., Nelson, J., & Ibarra, J. (2018). Cryptocurrency price prediction using tweet volumes and sentiment analysis. *SMU Data Science Review*, *1*(3), 1. https://scholar.smu.edu/datasciencereview/vol1/iss3/1

[31] Sakaki, T., Okazaki, M., & Matsuo, Y. (2010, April). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web* (pp. 851-860). https://doi.org/10.1145/1772690.1772777

[32] Chew, C., & Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PloS one*, *5*(11), e14118. https://doi.org/10.1371/journal.pone.0014118

[33] Jain, V. K., & Kumar, S. (2015). An effective approach to track levels of influenza-A (H1N1) pandemic in India using twitter. *Procedia Computer Science*, *70*, 801-807. https://doi.org/10.1016/j.procs.2015.10.120

[34] Ainin, S., Feizollah, A., Anuar, N. B., & Abdullah, N. A. (2020). Sentiment analyses of multilingual tweets on halal tourism. *Tourism Management Perspectives*, *34*, 100658. https://doi.org/10.1016/j.tmp.2020.100658

[35] Reyes-Menendez, A., Saura, J. R., & Filipe, F. (2020). Marketing challenges in the# MeToo era: Gaining business insights using an exploratory sentiment analysis. *Heliyon*, *6*(3). https://doi.org/10.1016/j.heliyon.2020.e03626

[36] Hassan, S. U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., ... & Herrera, F. (2020). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems*, *192*, 105383. https://doi.org/10.1016/j.knosys.2019.105383

[37] Al-Hashedi, A., Al-Fuhaidi, B., Mohsen, A. M., Ali, Y., Gamal Al-Kaf, H. A., Al-Sorori, W., & Maqtary, N. (2022). Ensemble classifiers for Arabic sentiment analysis of social network (Twitter data) towards COVID-19-related conspiracy theories. *Applied Computational Intelligence and Soft Computing*, *2022*, 1-10. https://doi.org/10.1155/2022/6614730

[38] Alenezi, M. N., & Alqenaei, Z. M. (2021). Machine learning in detecting covid-19 misinformation on twitter. *Future Internet*, *13*(10), 244. https://doi.org/10.3390/fi13100244

[39] Alabrah, A., Alawadh, H. M., Okon, O. D., Meraj, T., & Rauf, H. T. (2022). Gulf countries' citizens' acceptance of COVID-19 vaccines—A machine learning approach. *Mathematics*, *10*(3), 467. https://doi.org/10.3390/math10030467

[40] Fabian, P. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research 12*, 2825-2830. https://cir.nii.ac.jp/crid/1370005891170856713

[41] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text classification algorithms: A survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150

[42] Tyagi, A., & Sharma, N. (2018). Sentiment analysis using logistic regression and effective word score heuristic. *International Journal of Engineering and Technology (UAE)*, *7*(2), 20-23. https://www.researchgate.net/publication/325101249

[43] Jalal, N., Mehmood, A., Choi, G. S., & Ashraf, I. (2022). A novel improved random forest for text classification using feature ranking and optimal number of trees. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 2733-2742. https://doi.org/10.1016/j.jksuci.2022.03.012

[44] Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, *127*, 511-520. https://doi.org/10.1016/j.procs.2018.01.150

[45] Misra, S., Li, H., & He, J. (2020). Noninvasive fracture characterization based on the classification of sonic wave travel times.*Machine Learning for Subsurface Characterization,* 243-287.

[46] Bayhaqy, A., Sfenrianto, S., Nainggolan, K., & Kaburuan, E. R. (2018, October). Sentiment analysis about E-commerce from tweets using decision tree, K-nearest neighbor, and naïve bayes. In *2018 international conference on orange technologies (ICOT)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICOT.2018.8705796

[47] Vijayan, V. K., Bindu, K. R., & Parameswaran, L. (2017, September). A comprehensive study of text classification algorithms. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (pp. 1109-1113). IEEE. https://doi.org/10.1109/ICACCI.2017.8125990

[48] Buldin, I. D., & Ivanov, N. S. (2020, January). Text classification of illegal activities on onion sites. In *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)* (pp. 245-247). IEEE. https://doi.org/10.1109/EIConRus49466.2020.9039341

[49] Chan, L., & Martens, B. (2007). Openness in Digital Publishing: Awareness, Discovery and Access in *ELPUB*, 2007, 349-360.

[50] Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium* (pp. 1-7). https://doi.org/10.1145/2401603.2401605

[51] Sharma, A., & Dey, S. (2012, October). A comparative study of feature selection and machine learning techniques for sentiment analysis. In *Proceedings of the 2012 ACM research in applied computation symposium* (pp. 1-7). https://doi.org/10.1109/MLBDBI48998.2019.00062

[52] Hama Aziz, R. H., & Dimililer, N. (2021). SentiXGboost: enhanced sentiment analysis in social media posts with ensemble XGBoost classifier. *Journal of the Chinese Institute of Engineers*, *44*(6), 562-572. https://doi.org/10.1080/02533839.2021.1933598

[53] Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, *136*, 190-197. https://doi.org/10.1016/j.patrec.2020.05.035

[54] Wang, C., Deng, C., & Wang, S. (2020). Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost. *Pattern Recognition Letters*, *136*, 190-197. https://doi.org/10.1109/BigData.2017.8258507

[55] Erbek, F. S., Özkan, C., & Taberner, M. (2004). Comparison of maximum likelihood classification method with supervised artificial neural network algorithms for land use activities. *International journal of remote sensing*, *25*(9), 1733-1748. https://doi.org/10.1080/0143116031000150077

[56] Almaghrabi, M., & Chetty, G. (2020, October). Improving sentiment analysis in Arabic and English languages by using multi-layer perceptron model (MLP). In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)* (pp. 745-746). IEEE. https://doi.org/10.1109/DSAA49011.2020.00095

[57] Agarwal, S. (2013). Data mining: Data mining concepts and techniques. In *2013 international conference on machine intelligence and research advancement* (pp. 203-207). IEEE. https://doi.org/10.1109/ICMIRA.2013.45

[58] Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, *14*(1), 1-22. https://doi.org/10.1186/s13040-021-00244-z

[59] Lohar, P., Xie, G., Bendechache, M., Brennan, R., Celeste, E., Trestian, R., & Tal, I. (2021, August). Irish attitudes toward COVID tracker app & privacy: sentiment analysis on Twitter and survey data. In *Proceedings of the 16th International Conference on Availability, Reliability and Security* (pp. 1-8). https://doi.org/10.1145/3465481.3469193

[60] Yimam, S. M., Alemayehu, H. M., Ayele, A., & Biemann, C. (2020, December). Exploring amharic sentiment analysis from social media texts: Building annotation tools and classification models. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1048-1060). http://dx.doi.org/10.18653/v1/2020.coling-main.91

[61] Rupapara, V., Rustam, F., Shahzad, H. F., Mehmood, A., Ashraf, I., & Choi, G. S. (2021). Impact of SMOTE on imbalanced text features for toxic comments classification using RVVC model. *IEEE Access*, *9*, 78621-78634. https://doi.org/10.1109/ACCESS.2021.3083638

[62] Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF* IDF, LSI and multi-words for text classification. *Expert systems with applications*, *38*(3), 2758-2765. https://doi.org/10.1016/j.eswa.2010.08.066

[63] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1-13.. https://doi.org/10.1186/s12864-019-6413-7